Mots clés : Graphes. Vecteur propre ; matrices stochastiques ; matrices à coefficients positifs.

Le jury n'exige pas une compréhension exhaustive du texte. Vous êtes laissé(e) libre d'organiser votre discussion comme vous l'entendez. Il vous est conseillé de mettre en lumière vos connaissances à partir du fil conducteur constitué par le texte. Le jury demande que la discussion soit accompagnée d'exemples traités sur ordinateur. Il est souhaitable que vous organisiez votre présentation comme si le jury n'avait pas connaissance du texte. Le jury aura néanmoins le texte sous les yeux pendant votre exposé.

L'algorithme Pagerank du moteur de recherche Google

1. Introduction

Un moteur de recherche est un outil logiciel qui établit une liste (annuaire) des sites sur le web. Les internautes ont accès à cet annuaire en effectuant des requêtes sous forme de mots-clés. Son fonctionnement peut se décomposer en trois étapes :

1. L'*exploration* automatique du web pour récupérer les adresses des pages web à accès public : des agents appelés « robots » parcourent les sites à intervalles réguliers, en utilisant les liens hypertextes (lien permettant d'accéder à une page à partir d'une autre), ceci de façon automatique pour enregistrer de nouvelles adresses (URL).
2. L'*indexation* qui consiste à récupérer pour chaque page web des mots considérés comme significatifs. Ces mots sont enregistrés dans une base de données.
3. Le *classement* des pages web référencées dans la base de données afin que l'internaute qui effectue une recherche selon des mots clés recoivent en priorité les réponses selon un ordre de pertinence et d'importance.

Nous présentons ici l'algorithme "Pagerank" utilisé par le moteur de recherche Google pour l´étape de classement des pages web. Cet algorithme attribue à chaque page web une note qui va contribuer au classement des pages les plus importantes (celles qui ont la meilleure note) au moins importantes.

2. Modélisation(s).

Le web est représenté par un graphe orienté $G = (S, A)$ où S est un ensemble fini et A un sous-ensemble de $S \times S$:

- L'ensemble S, des sommets du graphe est en bijection avec l'ensemble des pages web référencées : Chaque page référencée est numérotée et l'ensemble S est l'ensemble des numéros des pages référencées soit un ensemble d'entiers de la forme $\{1, 2, ..., N\}$.

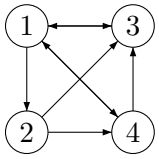

Figure 1: Un exemple de "web"avec 4 pages

- L'ensemble A des arêtes est en bijection avec l'ensemble des liens hypertextes : l'arête (i, j), souvent désignée par $i \to j$, est dans A si il existe sur la page i un lien pointant sur la page j.

Pour classer les pages selon leur importance (terme à définir), on souhaite leur attribuer à chacune une note, c'est à dire une fonction f de S dans $\mathbb{Q}^+$ déterminée en fonction des liens existants. L'idée de base est de considérer le web comme une démocratie, où chaque page est à la fois électeur et candidat, et un lien de la page i sur la page j représente un vote de l´electeur i vers le candidat j.

- Une première idée est d'affecter à chaque page le nombre de pages pointant sur elle (on compte les voix) ; Dans notre exemple, la fonction serait $f(1) = 2, f(2) = 1, f(3) = 3, f(4) = 2$. Le classement induit par un tel choix est peu fiable dans la mesure où la note d'une page peut-être aisément augmentée de façon artificielle par la création de pages pointant sur elle.

- Il parait raisonnable d'exiger qu'une page jugée importante transmette d'une certaine façon cette qualité aux pages sur lesquelles elle pointe. Une deuxième idée est donc d'affecter à une page la somme des notes des pages qui pointent sur elle. La fonction f serait ainsi définie comme solution du système :

$$\forall i \in S,\ f(i) = \sum_{j \to i} f(j).$$

  Il n'y a priori pas de raison que ce système ait une solution.

- Un autre défaut de la méthode précédente est qu'elle donne la même importance à chaque lien : deux pages d'égales importances n'ont pas le même pouvoir de vote, si l'une pointe sur une seule page alors que l'autre pointe sur de nombreuses pages, ce qui n'est pas équitable et encourage la vénalité. Le principe de la classification faite par Pagerank utilise l'idée précédente en limitant à 1 le vote de chaque élécteur (page) de façon uniforme. Si, pour chaque page j, n_j désigne le nombre de liens (vote) sur cette page la fonction f utilisée est définie par :

$$\forall i \in S, f(i) = \sum_{j \to i} \frac{f(j)}{n_j}.$$

  Sur notre exemple, toute solution est colinéaire à la solution $f(1) = 12, f(2) = 4, f(3) = 9, f(4) = 6$.

3. Une solution à ajuster

Le problème est de trouver une fonction de notation f qui convient et qui puisse être calculée. Le vecteur $(f(i))_i$ défini par la fonction f ci-dessus, doit être un vecteur propre relativement à la valeur propre 1 de la matrice

$$S = (\frac{1}{n_j} 1_{j \to i})_{i,j}.$$

Il faut donc s'assurer que la matrice S admet bien 1 comme valeur propre.
Remarque : Lorsque toute page est au moins pointée une fois, la matrice S vérifie les propriétés remarquables suivante:

- Les coefficients de S sont positifs,
- la somme des coefficients de chaque colonne de S est égale à 1.

La transposée d'une telle matrice est dite *matrice stochastique*.

Une matrice stochastique admet effectivement la valeur propre 1 et il en est donc de même pour sa transposée :

Proposition 1 *Soit P une matrice stochastique. Alors P admet 1 comme valeur propre associée au vecteur propre $(1, ..., 1)$.*

Corollaire 1 *Soit P une matrice stochastique. Alors ${}^tP$ admet 1 comme valeur propre.*

Dans le cas général, les pages qui ne sont pointées par aucune autre fournissent dans la matrice S des colonnes de 0. La matrice S est alors une matrice à coefficients positifs, majorée coefficient par coefficient par la transposée d'une matrice stochastique. Le théorème de Perron-Frobénius assure alors que son rayon spectral est plus petit que 1, atteint pour une valeur propre associée à un vecteur propre dont les coordonnées sont positives. Si ce rayon est < 1, il n'existe pas de fonction f (non triviale). Pour remédier à ce problème, on peut remplacer ces colonnes nulles par des colonnes de $\frac{1}{N}$ par exemple.

Une autre difficulté provient du caractère *réductible* de la matrice S :

Définition : Une matrice S est dite réductible si il existe une matrice de permutation Σ et des matrices A, B, C telles que :

$$S = \Sigma \begin{pmatrix} A & B \\ 0 & C \end{pmatrix} \Sigma^{-1}.$$

Pour la matrice S, être irréductible signifierait : de toute page web, on peut atteindre n'importe quelle page web en utilisant uniquement des liens hypertextes. Cette hypothèse est peu vraisemblable et contredite par des études statistiques faites sur le web.
Il est pourtant important d'utiliser une matrice irréductible pour le calcul de f. La proposition ci-dessous montre que sous cette hypothèse, il existe une vecteur qui convient, uniquement déterminé si on ajoute une condition de normalisation (par exemple, $\sum_i f(i) = 1$).

Proposition 2 *Soit P une matrice stochastique irréductible. Le sous-espace propre* $\ker(P - Id)$ *est de dimension* 1*, engendré par un vecteur dont les coordonnées sont strictement positives.*

La matrice S est alors remplacée par la matrice $T := (1-\alpha)S + \alpha J$ où J est une matrice dite de pertubation. Elle représente la possibilité pour l'internaute d'écrire une URL au hasard plutot que de naviguer uniquement avec des liens. Ici, la matrice J est la matrice (N, N) dont tous les coefficients sont égaux $\frac{1}{N}$. Le nombre réel α est dans l'intervalle $[0, 1]$ (dans l'algorithme, la valeur $\alpha = 0,15$ est utilisée).

4. Le calcul du vecteur propre.

Il y a plusieurs billions de pages web. Le vecteur propre à calculer est donc solution d'un système linéaire de taille plusieurs billions. Une méthode pour le calculer est de l'approcher :
On choisit un vecteur x_0 à coordonnées strictement positives, puis on définit la suite $x_{k+1} = Tx_k$.

Proposition 3 *La suite normalisée* $(x_k/\|x_k\|_1)$ *converge vers un vecteur propre de T, relatif à la valeur propre* 1 *et dont les coordonnées sont toutes strictement positives.*

La limite de cette suite est donc la solution cherchée. Etant donnée la taille des objets à manipuler pour faire le calcul, même approché, l'efficacité du calcul est un enjeu crucial. Le calcul du vecteur propre prend plusieurs jours et doit etre actualisé régulièrement.

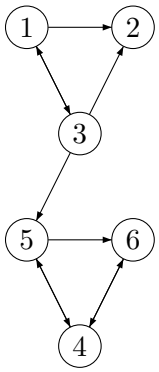

Figure 2: Un exemple de "web"avec 6 pages

Suggestions de développements.

Soulignons qu'il s'agit d'un menu à la carte et que vous pourrez choisir d'étudier certains points, pas tous, pas nécessairement dans l'ordre et d'une façon plus ou moins fouillée. Vous pouvez aussi vous poser d'autres questions que celles indiquées plus bas. Il est très vivement conseillé que vos investigations comportent une partie traitée sur ordinateur et, si possible, des représentations graphiques de vos résultats.

- On pourra démontrer les diverses propositions.
- On pourra étudier la vitesse de convergence de la suite $(x_k)_k$.
- On pourra expliciter les différents tentatives de définition d'une fonction de notation et leurs significations sur un petit exemple, par exemple celui de la figure 2.
- On pourra proposer une interprétation probabiliste des matrices S, T et du vecteur limite de la suite $(x_k/\|x_k\|_1)$.